

AVEC: An Assessment of Visual Encoding Ability in Visualization Construction

Lily W. Ge

Computer Science

Northwestern University

Evanston, Illinois, USA

wanqian.ge@northwestern.edu

Yuan Cui

Computer Science

Northwestern University

Evanston, Illinois, USA

yuancui2025@u.northwestern.edu

Matthew Kay

Computer Science and

Communication Studies

Northwestern University

Chicago, Illinois, USA

mjskay@northwestern.edu

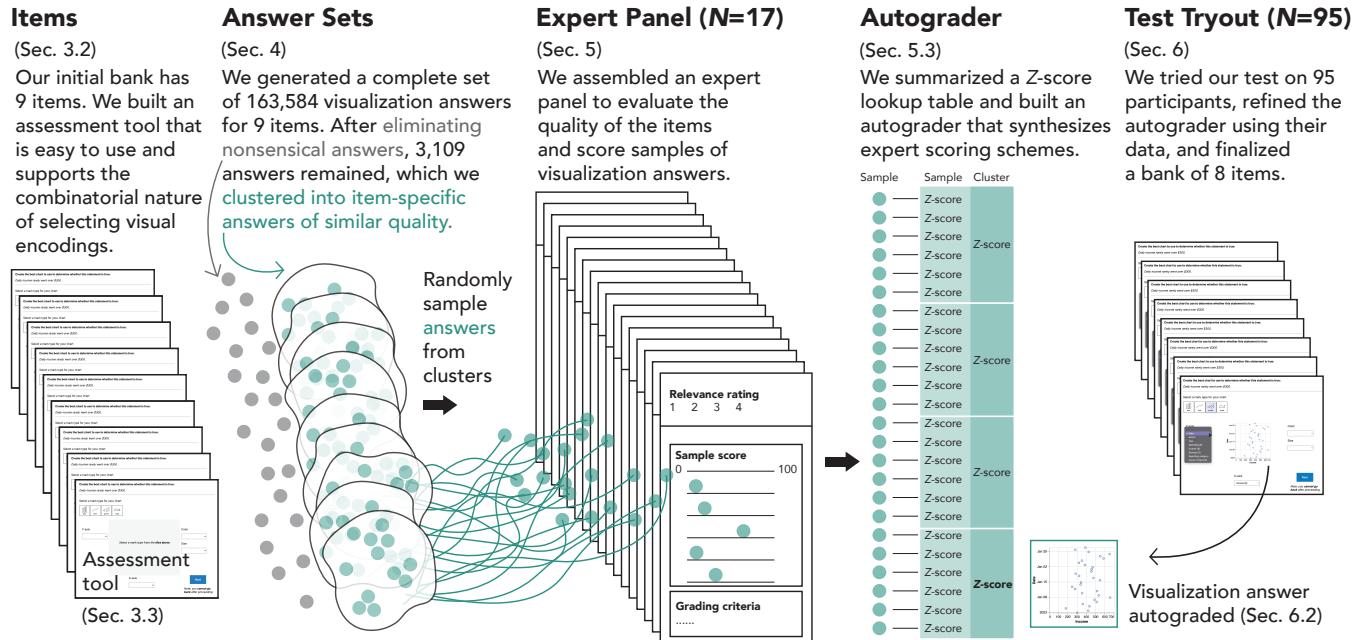


Figure 1: The systematic development process of AVEC. We created a design space to generate a bank of 9 items (Section 3.2) and built an assessment tool that is easy to use and supports the combinatorial nature of selecting appropriate visual encodings (Section 3.3). From a complete set of possible visualization answers per item (Section 4), we clustered visualization answers of similar quality after first eliminating poor quality answers. We assembled an expert panel to rate the quality of the items and samples from the clusters of remaining visualization answers (Section 5). We then built an autograder that synthesizes expert scoring schemes using expert ratings and data from test tryout (Section 5.3 and Section 6.2), which can automatically assign scores to visualization answers.

Abstract

Visualization literacy is the ability to both interpret and construct visualizations. Yet existing assessments focus solely on visualization interpretation. A lack of construction-related measurements hinders efforts in understanding and improving literacy in visualizations. We design and develop AVEC, an assessment of a person's *visual encoding ability*—a core component of the larger process of visualization construction—by: (1) creating an initial item bank using a design space of visualization tasks and chart types, (2) designing an assessment tool to support the combinatorial nature of

selecting appropriate visual encodings, (3) building an autograder from expert scores of answers to our items, and (4) refining and validating the item bank and autograder through an analysis of test tryout data with 95 participants and feedback from the expert panel. We discuss recommendations for using AVEC, potential alternative scoring strategies, and the challenges in assessing higher-level visualization skills using constructed-response tests. Supplemental materials are available at: <https://osf.io/hg7kx/>.

CCS Concepts

- Human-centered computing → Empirical studies in visualization; Empirical studies in HCI.

Keywords

Visualization literacy, Visualization construction, Measurement

ACM Reference Format:

Lily W. Ge, Yuan Cui, and Matthew Kay. . AVEC: An Assessment of Visual Encoding Ability in Visualization Construction. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3706598.3713364>

1 Introduction

The ability to construct effective visualizations is a critical component of *visualization literacy* [4, 11], analogous to the way in which writing is inseparable from textual literacy. With the rapid increase in data availability and accessibility, visualizations are not only created by professional designers [22, 43, 49], but also by members of the general public; for example, people who are not in data-oriented jobs or who might not be trained in visualization are nevertheless constructing visualizations at work (e.g., through Excel) or in a personal context (e.g., through personal visualization and visual analytics [28]). Despite increased participation in constructing visualizations and efforts to support non-experts [29, 51], the current focus of visualization literacy studies and assessments is still on visualization interpretation [9, 20, 24, 32, 41]. We lack valid measurements of construction-related abilities, which we need to effectively measure, teach, and improve these skills in novices and amongst the general public.

Constructing visualizations, however, is a complex and multi-step process [13, 31], which includes transforming raw data into a form suitable for visualization (e.g., cleaning, filtering, and aggregating as necessary [52]), translating data variables to visual variables (e.g., setting the position of a data point by mapping data to desired x and y axes [36]), and transforming the visual encodings to convey a meaningful message (e.g., applying appropriate scaling functions or applying creative visual metaphors [56] to facilitate interpretation). All of these components make up the broad notion of a *visualization construction ability*.

Although there are many steps involved in visualization construction, one of its core components is selecting appropriate *visual mappings* [13]. We designate a person's *ability to appropriately map data to visual channel(s) for effectively answering data-relevant questions* their **visual encoding ability**. This ability, which is governed by established visualization principles [13, 36] and rules (e.g., the Grammar of Graphics [53]) and common to all visualization design, is critical for constructing well-formed visualizations and for visualization creators to master. We therefore focus on the visual encoding ability as a first step in systematically assessing construction-related abilities.

The complexity of visualization construction also translates into challenges in measuring construction-related abilities. The traditional multiple-choice question format cannot accurately represent the design space a visualization creator has to work with. Visually encoding data requires a designer to appropriately apply grammatical rules of visualization design and consider multiple data variables and visual encodings at once; the design space quickly explodes to thousands of possible combinations of data mappings.

This *combinatorial* nature is inherent to selecting appropriate visual encodings and must also be present in measurements of such abilities.

To create an assessment of the visual encoding ability in visualization construction, we:

- (1) **Created 9 initial items to measure the visual encoding ability.** To create the items, we first constructed an item design space consisting of 3 visualization tasks and 8 chart types. We used it to generate 38 candidate constructed-response items, each of which asks the test-taker to create a visualization for answering a true-or-false (T/F) statement about a dataset. We selected the 9 items that make up the initial item bank by varying the number of necessary data variables, mark types, and size of datasets.
- (2) **Developed a web-based assessment tool that can support the combinatorial nature of selecting appropriate visual encodings.** This tool allows test-takers to create a wide variety of visualization answers for the items, while still being **easy to use** without requiring prior knowledge of a specific visualization tool or programming language. This ensures that differences in test-takers' scores are due to differences in their abilities and not the usability of the tool, which contributes to the validity of the assessment.
- (3) **Created an autograder based on expert scores of visualization answers to our assessment items.** We invited an expert panel of 17 Information Visualization researchers to assess the content validity of candidate items and provide grading criteria for clusters of similar answers to the items.
- (4) **Conducted test tryout to assess the quality of our items and to refine the autograder.** Ninety-five online participants used our web-based assessment tool to answer the 9 candidate items in a test tryout study. We used Item Response Theory (IRT) analysis to measure how easy or difficult each item is and how reliably each item separates people of different abilities. We also used the test tryout data to refine the autograder and assign final grades to participants' answers.
- (5) **Refined the bank to a final set of 8 constructed-response items, which have high content validity, have varying difficulty, and can differentiate people of different levels of ability.** We finalized the item bank of AVEC using feedback from the expert panel and results of the IRT analysis, removing one item with low content validity and high variance in experts' scores.

Reflecting on our efforts to go beyond multiple-choice formats, AVEC demonstrates the importance of and the complexities involved in measuring higher-level abilities in visualization literacy. Dealing with these complexities necessitates a line of future research that could transform both the ways that we assess visualization skills and our understanding of visualization literacy.

2 Related Work

2.1 Visualization Construction

Visualization, as described by Card et al. [13], is the mapping of data to a visual form that supports human sense-making. A sequence of steps are involved in this process, including data transformations,

visual mappings, and view transformation, and have been referred to by Card et al. as a *reference model*. This information visualization reference model has been extended, adapted, and refined in prior work [14, 16, 31, 50]. We focus on the core of this model, which is the **visual mapping** transformation: the mapping from data to visual representation. Visual mappings involve encoding data variables to appropriate visual structures, such as visual marks (e.g., point, line) and visual channels (e.g., position, color, size). The mapping of data to visual structures are governed by established principles such as the ranking of perceptual tasks [36], effectiveness and expressiveness criteria [36], and the Grammar of Graphics [53].

Visualization construction is a complex process, and previous work has studied how people create visualizations [30] and how to improve the ability to create visualizations. For instance, both Alper et al. [3] and Gäbler et al. [23] developed interventions to help young children learn about bar charts; He and Adar [27] created VizItCards, a toolkit to facilitate collaborative visualization design in classroom instruction; Bishop et al. [8] developed Construct-A-Vis, a tool that scaffolds the visual mapping process and supports children in free-form visualization creation; and Adar and Lee-Robbins [1] created Roboviz, a game-based activity for visualization courses to facilitate the creation of interactive visualizations. However, limited work has focused on *measuring* how well people can construct visualizations, which requires valid measurements of construction-related abilities. Such measurements can help identify areas of improvement in novice learners and evaluate the effectiveness of proposed interventions.

2.2 Assessments of Visualization Literacy

Researchers have developed several assessments and frameworks for measuring and studying visualization literacy. Boy et al. [9] demonstrated a method for assessing visualization literacy and designed tests for line charts, bar charts, and scatterplots. Lee et al. [32] developed the Visualization Literacy Assessment Test (VLAT) that contains 53 multiple-choice items that measures the ability to “read and interpret visually represented data in and to extract information from data visualizations”. Following VLAT, Pandey and Ottley [41] developed Mini-VLAT to more efficiently measure this ability with fewer items. Börner et al. proposed a data visualization literacy framework (DVL-FW) that aims to guide the teaching and assessment of visualization literacy, which includes both visualization interpretation and construction. Extending prior definitions of visualization literacy that focus on the interpretation of well-formed charts, Ge et al. [24] developed the Critical Thinking Assessment for Literacy in Visualizations (CALVI) composed of a 45-item bank to assess people’s ability to “read, interpret, and reason about erroneous or potentially misleading visualizations”. Cui et al. [20] then applied Computerized Adaptive Testing (CAT) methods to develop A-VLAT and A-CALVI to adaptively assess a person’s basic visualization interpretation ability and critical thinking ability in detecting visualization misinformation.

However, visualization literacy consists of the ability to both interpret and construct visualizations [4, 11]. Yet many of the existing assessments focus on interpretation alone. Adelberger et al. [2] also noted this lack of measurement of visualization construction abilities and created a set of multiple-choice items to

evaluate Iguanodon, a gamified intervention they designed to improve visualization literacy. However, their items only test for the overplotting issue and do not cover the core of visualization construction. Valid measurements of construction-related abilities can help evaluate the effectiveness of targeted interventions to improve such abilities and support research efforts in the holistic study of visualization literacy. The need of such measurements was also called for in a CHI 2024 workshop: *Toward a More Comprehensive Understanding of Visualization Literacy* [26].

2.3 Item Formats Beyond Multiple Choice

Existing assessments of visualization literacy often use multiple-choice items [9, 19, 20, 24, 32, 41]. However, simply choosing from a short list of answers is not an accurate representation of construction-related abilities, and thus is not a format that can sufficiently support measuring the ability to appropriately map data to visual channel(s). Constructed-response items, on the other hand, require a test-taker to “generate an answer rather than select from a short list of options” [6]. They are often used to measure higher-level and more complex skills in literature, mathematics, and music [34]. While these items can better capture the skill of interest, they are harder to grade. The Educational Testing Service (ETS), for example, recognizes this and has applied automated scoring to support the grading of the constructed-response items [34]. In the context of a visualization assessment, we might opt to use a computer-mediated constructed-response item format, rather than purely asking test-takers to draw visualizations by hand, for example. Allowing test-takers to directly control the mapping of data variables to visual encodings serves as a proxy to hand-drawing visualizations and also preserves the combinatorial nature of selecting appropriate visual encodings (more details in Section 3.3).

3 Test Development: Items and Assessment Tool

Visual mappings are a core component of both the visualization reference model [13] and visualization grammar [53], making them essential for visualization creators to master. We dub the *ability to appropriately map data to visual channel(s) for effectively answering data-relevant questions* a person’s *visual encoding ability*, and focus on developing an assessment that measures this ability. We aim to design the assessment for non-experts, including members of the general public and students who are beginning to learn how to construct visualizations.

We use constructed-response items in our assessment because generating an answer—rather than choosing from a short list of answers—aligns better with the combinatorial nature of mapping data to visual channels. An *item* of AVEC consists of a constructed-response question that asks the test-taker to create a chart for answering a true-or-false (T/F) statement about a dataset. The format of each item is as follows:

“Create the best chart to use to determine whether this statement is true: [T/F statement].”

Thus, for a diverse bank of items, we need a set of T/F statements whose reasonable visualization answers cover a diverse set of commonly-used visualization tasks and chart types (Section 3.1 and Section 3.2). Additionally, the items must be administered via

Table 1: The data variables in each dataset. The size of the dataset for each item may be different, but each item asks about a subset of these variables.

Data variable	Description
Date	A recorded date (year-month-date) to track daily spending, income, or savings.
Month	The month.
Year	The year.
Income	The amount of income (\$) on each recorded date.
Savings	The amount of savings (\$) on each recorded date.
Spending	The amount of spending (\$) on each recorded date. In this hypothetical dataset, the person spends money on only one spending category each day.
Spending category	Daily spending category: Entertainment, Food, Transport, or Utilities.

an easy-to-use assessment tool that also supports the combinatorial nature of selecting appropriate visual encodings (Section 3.3).

3.1 Item Design Space

To systematically generate a set of suitable candidate T/F statements that will allow for a diverse set of visualization answers, we created a design space composed of a variety of visualization tasks and chart types.

Visualization Tasks. We started with the initial set of 8 tasks from VLAT [32], which was designed to assess the visualization interpretation ability of the general public. We **excluded 2 tasks that are lower-level (involving local point judgements)** to focus on higher-level tasks (involving visual patterns with many data points), because the visual encoding ability that we intend to measure involves whether people can create an appropriate *visual form* to answer a data-relevant question. Specifically, we excluded *Retrieve Value* and *Find Extremum* because they are single-point judgements, so we consider them lower-level. Additionally, we aimed to create a diverse set of items that would lead test-takers to construct a variety of visualization types. To this end, we **merged tasks that would generally lead to similar visualizations**. Namely, we merged *Determine Range*, *Find Anomalies*, and *Find Clusters* into *Describe Distributions*.¹ We kept the remaining 2 tasks from the original set of 8 as-is (i.e., *Find Trends / Correlations* and *Make Pattern Comparisons*) because they are higher-level tasks and are distinct from the task of describing distributions. As a result, our reduced set contains 3 tasks:

- *Describe Distributions*: describe the spread of the values of a variable and/or how frequently different values occur.
- *Find Trends / Correlations*: find the relationship between two variables or how a variable changes over time.
- *Make Pattern Comparisons*: compare patterns (e.g., data distributions, trends, or relationships) in the dataset.

Chart Types. To identify a set of chart types that the visualization answers should cover, we started with a set of 12 chart types from VLAT [32]. We aimed to compile a diverse set of chart types

that cover our set of 3 tasks and can support varying the difficulty levels of items in the bank. However, for usability purposes, we also wanted to keep our assessment tool’s interface consistent across items so that test-takers do not have to learn different interfaces for different items—in order to reduce the influence of the interface on a test-taker’s performance. Out of the 12 initial chart types, a majority of chart types have *x* and *y* axes, which could be kept consistent across items. Thus, we retained 8 Cartesian chart types,² as this would allow us to **have a diverse set of chart types while maintaining a consistent chart-creation interface**. Our resulting set contains 8 total chart types,³ and each supports a maximum of 4 encoding channels (i.e., *x-axis*, *y-axis*, color, size).

3.2 Item Generation

We generated one T/F statement for each candidate item using the item design space. For all T/F statements, we used a dataset in the context of personal finance, because this topic is relevant to most people. We generated datasets of varying sizes (e.g., a year’s or a month’s data) to introduce some diversity between items, but all of the datasets contain the same set of data variables, as described in Table 1. Each T/F statement asks about a subset of the variables in this dataset. This reduced the need for test-takers to learn new datasets during the assessment: how well one can get familiar with a new dataset is not necessarily part of our measure of the ability to map data to visual channels.

We first generated T/F statements with a differing number of *necessary data variables* (i.e., variables that must be visualized to answer the T/F question correctly) for each combination of visualization task and chart type. This was our attempt at generating items with varying difficulty. The ultimate difficulty of these items will be determined by IRT analysis (Section 6.5). For 6 of the 8 chart types, we generated two candidate T/F statements per combination of task and chart type, resulting in 36 items (2 items × 6 chart types × 3 visualization tasks). For stacked bar chart and stacked area chart, we generated one candidate T/F statement each (under the task of *Make Pattern Comparisons*), because they require three and only

¹We renamed the original *Characterize Distribution* from VLAT to better reflect this merged set, which includes tasks related to describing the spread of data values rather than only characterizing distributions.

²Between stacked bar chart and 100% stacked bar chart, we only kept stacked bar chart. Because of their similarity, keeping both would add little value to having a diverse set.

³The 8 chart types are bar chart, stacked bar chart, histogram, line chart, scatterplot, bubble chart, area chart, and stacked area chart.

Task	ID	True-or-false statements	Mark Types
Describe Distributions	DD.1	Daily income rarely went over \$300.	
	DD.2	Total monthly utilities spending was generally lower than \$300.	
	DD.3	Total yearly spending rarely exceeded \$10,000.	
Find Trends Correlations	FTC.1	Daily spending often fluctuated between about \$10 to about \$30.	
	FTC.2	Daily savings was generally higher when income was higher.	
	FTC.3	Total monthly utilities spending strictly decreased from Jul to Oct.	
Make Pattern Comparisons	MPC.1	The range of daily transport spending was wider than food spending.	
	MPC.2	The number of times that daily spending was around \$50 increased from 2020 to 2023.	
	MPC.3	Daily food spending consistently decreased at the same rate as transport spending.	

Figure 2: The set of T/F statements in our initial item bank. Each statement has its own set of mark types that we think are most relevant based on its relevant data type(s) and task, which covers a variety of chart types.

three necessary variables. Thus, we created $36 + 2 = 38$ candidate T/F statements. We referred to a summarization of suitable chart types for different data types by authors of VLAT [32] and inferred the set of most relevant mark types for each T/F statement. For instance, scatterplots are less likely to be useful for items with T/F statements that require yearly aggregations, and line charts are more likely to be useful for T/F statements that require order in the data (e.g., trend over time).

We then reviewed the set of candidate T/F statements and noticed redundancies that involved the same type of statement that only differ in the quantitative variable of interest (e.g., “[quantitative variable] rarely went over [numeric value as threshold]”). To remove redundancies and ensure a variety of items, we selected T/F statements from each visualization task category while ensuring a diverse set of relevant chart types,⁴ a varying number of necessary variables, and a varying size of datasets. This allowed us to arrive at a set of T/F statements that can lead to a diverse set of visualization answers that also cover all 3 visualization tasks. The result is a set of 9 items, with 3 items per task, as shown in Figure 2.

3.3 Assessment Tool

We need an assessment tool that is easy to use, able to support the combinatorial nature of selecting appropriate visual encodings, and expressive enough to allow test-takers to construct a good answer to the items. This led to the following design considerations:

- **(DC.1) Does not require coding expertise:** The target test population for AVEC includes non-experts in visualization, so we cannot assume they have any programming experience.

⁴We did not end up including T/F statements from bubble chart combinations due to a conflicting property of such T/F statements: the statements tend to be overly long and difficult to parse, but the visualization answer always includes all 3 of the quantitative variables in the dataset, which can make it easy to arrive at the answer.

- **(DC.2) Easy to use:** Our goal is to assess visualization-related abilities, not expertise with the interface—therefore we need an assessment tool that non-experts can pick up quickly with little training. This will help ensure that differences in test-takers’ scores are due to differences in their abilities and not the usability of the tool.
- **(DC.3) Familiarity of user interface:** Similar to the previous consideration, we do not want our measurement confounded by a test-taker’s familiarity with a specific visualization tool—therefore, we require an interface that is familiar to a broad set of users (e.g., one which uses standard UI widgets).
- **(DC.4) Consistency across items:** An assessment interface that is consistent across items reduces the need for the test-taker to learn new parts of the interface during the test, reducing noise in our measurement.
- **(DC.5) Expressive:** Even though we need to keep the assessment interface simple, it must be able to support the construction of a wide variety of chart types to give test-takers the flexibility to construct visualization answers of varying quality.

We went through an iterative design process to develop an assessment tool that meets all of the above design considerations. To broaden the user group for our assessment, we eliminated the need for programming experience by keeping controls at a high level: users only need to decide which visual channels they want to map data variables to **(DC.1)**. Behind the scenes, we manipulate Vega-Lite [46] specifications according to the test-taker’s desired mappings. The color schemes we used are color blind safe.⁵

⁵See supplemental materials for the results from Adobe’s color accessibility checker.

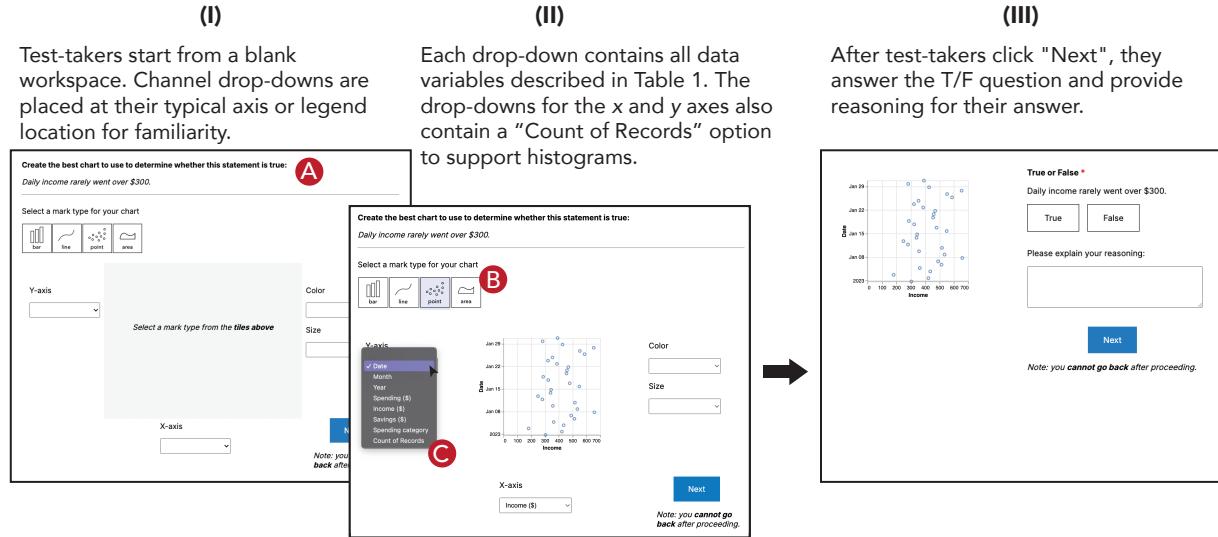


Figure 3: The interface of the assessment tool, which displays the item's question text (A) and includes an interactive area for constructing the visualization answer that allows test-takers to select their desired mark type (B) and contains drop-downs (C) to control the mapping between data to visual channels. On a separate page after a test-taker constructs their answer, they are asked to select an answer for the T/F question.

Early versions⁶ of the interface included individual controls in a drag-and-drop format (similar to visualization tools like Tableau) for different data transformations (e.g., binning, sorting, mean, sum, count), scale transformations (e.g., reversing or truncating scales), and a larger set of encoding choices (e.g., sequential, diverging, qualitative colors). However, when piloting the interface, we found that the drag-and-drop features were too overwhelming for some users to navigate and were difficult to learn through short training. As our goal was to create an interface that did not require extensive training (**DC.2**) and which did not simply measure a user's familiarity with some particular visualization tool (**DC.3**), we changed the interface to use more typical drop-down widgets and refined our scope to focus on visual mappings, the core component of the visualization reference model [13]. We refined the tool to support a maximum of 4 encodings (x-axis, y-axis, color, and size) and 4 mark types, which together determine the resulting chart type. The 4 mark types are *bar* (for bar chart, stacked bar chart, histogram), *line* (for line chart), *point* (for scatterplot, bubble chart), and *area* (for area chart, stacked area chart) (**DC.5**, Figure 3.B).

In the refined version of the assessment tool, we automatically apply aggregation functions for the purpose of supporting the creation of all of the chart types in our design space. In cases where the desired chart type requires aggregation, we manipulate Vega-Lite specifications to apply aggregation depending on the selection of mark types and data variable types. We included 2 aggregation functions:

- *count* is included to support histograms. If a test-taker selects the *Count of Records* option for one of the positional axes, then we apply the count aggregation and bin the variable on the other axis.

⁶See supplemental materials for details of early iterations.

- *sum* is included to support stacked bar and stacked area charts. If a test-taker selects a *bar* or *area* mark, maps a categorical variable to one of the positional axes and/or color and/or size, and maps a quantitative variable to the other axis, then we automatically sum the quantitative variable within combinations of the categorical variables.

Having a minimum set of built-in aggregation functions simplified the interface compared to earlier prototypes, which required test-takers to manually specify aggregation functions, but which participants found difficult to learn in pilots (**DC.2**). We further discuss the importance of the role of the assessment tool in measurements of complex construction-related skills in Section 7.5.

Throughout the assessment, we display each item in the same format for consistency (**DC.4**, Figure 3.A): the question text followed by an interactive area for the test-taker to construct their visualization answer (Figure 3). We placed the drop-downs close to the channel they control (or, for color and size, where their legends might often appear on simple charts), taking advantage of familiar spatial associations (**DC.3**, Figure 3.C). On a separate page, after the test-taker constructs the chart, we ask them to select the answer to the T/F question and provide their reasoning (Figure 3.III). We included this for us (and future test administrators) to use as a resource to revise the T/F statements if needed, but the test-takers' answers to the T/F questions do not affect the assessment of their visual encoding ability.

4 Test Development: Visualization Answers

The combinatorial nature of selecting appropriate visual encodings implies there should be a large number of possible answers of varying quality for each item in our assessment. Unlike the traditional multiple-choice format—where the correctness of the answer is

binary—we need a grading approach that allows test administrators to evaluate the quality of a broad spectrum of possible visualization answers.

It would be time-consuming to manually grade each possible visualization answer our assessment tool can generate. Therefore, we built an autograder for each item. Specifically, we (1) generated the complete set of possible visualization answers (as Vega-Lite specifications) to each question (Section 4.1); (2) used established rules and design principles from the literature [15, 36] to eliminate nonsensical answers, which will receive the lowest score (Section 4.2); (3) grouped the remaining answers into clusters we believe might receive similar scores from experts (Section 4.3); (4) randomly selected visualization answers from each cluster to form representative sets of visualization answers for each item. These representative sets of answers will be scored by an expert panel (Section 5). These expert scores will allow us to refine clusters (if needed) and will serve as the basis for our autograder.

4.1 Complete Sets of Visualization Answers

Our assessment tool, as explained in Section 3.3, gives test-takers the flexibility to map 7 data variables to 4 encoding channels and choose among 4 mark types. The test-taker manipulates each encoding channel through a drop-down list and is allowed to leave any channel “empty”. Thus, with the “empty” option, there are (4 mark types \times 8 options on x -axis \times 8 options on y -axis \times 8 options on color \times 8 options on size) 16,384 combinations. In addition to data variables, test-takers can also map *Count of Records* to one positional axis and a data variable to another to create histograms. This generates an additional (4 mark types \times 7 data variables on x -axis \times 8 options on color \times 8 options on size) 1,792 combinations,⁷ yielding $16,384 + 1,792 = 18,176$ possible visualization answers for each item.

4.2 Initial Elimination

It would be impractical to ask the expert panel to rate the complete set of 163,584 visualization answers (18,176 answers for each of the 9 items), so we reduced the set of visualization answers for each item using the following criteria:

- **Whether all of the necessary data variables for that item are in the visualization specification:** Visualizations that do not contain all of the variables asked about in the T/F question statement would not be able answer that item, so we eliminate these answers.
- **Whether the visualization passes the linter rules:** We reviewed linter rules from VizLinter [15] and only included the rules that pertain to the functionalities of our assessment tool. For example, the rule of *not using both bin and aggregate on the data at the same time* was excluded because our assessment tool automatically bins or aggregates depending on the data variable mapped, so it would not be possible to violate this rule. As a result, we curated the following three rules to use for this initial elimination:

⁷We fixed *Count of Records* to the y -axis when generating the combinations because versions with *Count of Records* on the x -axis would be redundancies that add little value to summarizing the grading rubric, which were combinations we could eliminate immediately to reduce the answers under consideration.

- Use different fields for x -axis and y -axis
- Use no more than one continuous data in the x and y channels for mark ‘bar’ and ‘tick’
- Mark ‘bar’, ‘tick’, ‘line’, ‘area’ require some continuous variable on x or y

Additionally, the assessment tool allows for some combinations of mark type and data variable mappings that would lead to visualizations that make it *theoretically impossible* to answer the corresponding T/F question; we eliminated these answers. For example, visualization answers that use the *bar* or *area* mark and contain *Month* or *Year* were eliminated for items that ask about correlation between daily values of two variables (FTC.2), as aggregating by *Month* or *Year* would not allow the test-taker to judge daily correlation (see supplemental materials for a representative *theoretically impossible* answer for each rule). The visualization answers eliminated during this phase are of poor quality.

In total, 160,475 visualization answers failed the initial elimination criteria, leaving 3,109 answers that must be graded (see Figure 4 for a breakdown by item).

4.3 Answer Clusters

It would still be impractical to have experts rate all 3,109 remaining non-eliminated answers. Instead, we first created clusters of visualization answers that we consider to be similar to each other, have experts score answers sampled from those clusters, then build our autograder based on those scores.

To build the clusters, we first partitioned the visualization answers to each item into broad clusters using high-level rules, such as *Do core quantitative variable(s) use positional encodings?*. We then built a tool to display all of the visualization answers in a given cluster, which allowed us to quickly look through clusters and judge their quality and similarity. The authors reviewed all clusters independently, then iteratively refined the clusters and high-level clustering rules, reaching consensus after regular discussion. The resulting high-level rules closely aligned with prior work on the ranking of perceptual tasks [36].

Next, we split the initial broad clusters into smaller clusters based on our judgments of the effectiveness and expressiveness of the visualizations. For example, a scatterplot with both necessary quantitative variables on positional axes is more effective for judging correlation than a plot with both variables mapped to color and size channels (FTC.2). Thus, we split the broad clusters based on properties of answers that could impact the visual form of the visualization: mark type, what data type is mapped to the encoding channels, and (for stacked bar or area) the orientation of the chart (orientation of a stacked bar can affect whether a subgroup of interest is aligned to the baseline, and therefore the difficulty of estimating values from that subgroup).

While forming the *clusters*, we further identified visualization answers that make it *practically impossible* or extremely difficult to answer the question statement and eliminated them based on item-specific rules. For example, it would be practically impossible to determine the correlation between two quantitative variables if the mark type is *bar* and one of the necessary quantitative variables is mapped to the *size* channel instead of the positional axis (FTC.2).

QID	Eliminated	Clusters (no particular order)
DD.1	16,921	24 4 30 30 60 60 72 6 52 14 240 256 78 128 2 40 10 18 54 50 6 12 9
DD.2	18,144	2 9 9 4 2 2 2 1 1
DD.3	17,940	20 1 1 28 2 18 1 1 112 10 22 1 1 2 1 1 4 1 1 6 2
FTC.1	17,880	72 6 14 78 2 40 10 18 50 6
FTC.2	17,648	4 204 60 4 64 64 78 50
FTC.3	18,126	18 2 9 9 4 2 2 2 1 1
MPC.1	17,785	12 24 22 13 54 60 22 96 30 6 2 6 8 32 2 2
MPC.2	17,931	28 2 10 14 2 28 30 110 2 18 1
MPC.3	18,100	20 8 20 8 2 6 10 2

Figure 4: The left-most column in gray indicates the number of answers that were initially eliminated for each item. The set of numbers in colored boxes indicates the number of visualization answers that were in each item-specific cluster. The clusters are unordered; we use the expert scores to determine the relative order after conducting the expert panel (Section 5).

The resulting *clusters* for each item represent sets of visualization answers that we would expect to receive similar scores. However, we wanted to both validate our clustering and receive broader expert input to determine final scores. Therefore, we formed an expert panel to collect experts' scores on answers sampled from these clusters.

5 Expert Panel

The goal of the expert panel is threefold: (1) collect expert ratings of relevance to compute the content validity index (cvr) of each item (cvr is a measure of how well an item on a test measures what it is intended to measure [42]); (2) gather expert criteria for assessing the quality of the visualization answers in order to validate our clustering scheme; and (3) use experts' scores of visualization answers sampled from our clusters to build autograders. We recruited 17 researchers in Information Visualization who have obtained ($N = 13$) or are obtaining ($N = 4$) a doctorate.

The expert panel was conducted on Qualtrics. For each item, we generated 5 representative sets of visualization answers. This was to have the expert panel rate a variety of samples. Each representative set contained one answer from each of its item's clusters, selected at random (see Section 4.3 and the breakdown of clusters by item in Figure 4). Each expert was randomly assigned 3 items from the set of 9, then shown one representative set of answers for each item they were assigned. We ensured that each item was seen by at least 5 experts, which is important for the purpose of computing the content validity index.

We asked each expert to (1) rate the relevance of each question statement to assessing a test taker's ability to create a visualization that serves that item's corresponding task, (2) score each visualization answer on a scale from 0 (poorly designed and unable to answer the question) to 100 (well designed and able to answer the question); this yields a holistic judgment of answer quality, which we later use to build our autograder, and (3) explain their grading criteria.

5.1 Content Validity of AVEC

The content validity index (cvr) is a metric that uses expert ratings to assess how well test items measure what they are intended to measure. The cvr is typically calculated from relevance ratings on a 4-point scale and is the proportion of experts who rated an item 3 (quite relevant) or 4 (highly relevant) [42]. Items with a cvr below 78% are candidates for revision [42]. We use cvr in conjunction with IRT analysis in Section 6.6 to finalize the item bank.

In our expert panel, if an expert rated the relevance of an item below 3, they were asked to provide free-text comments to explain their reasoning. We used their comments to evaluate whether the revision for the item would be minor; if minor, the item would not require a second round of expert review [42]. Two items were candidates for revision:

- DD.3: "Total yearly spending did not exceed \$10,000 in 2020" (cvr: 0.33, task: *Describe Distributions*). Several experts mentioned that DD.3 only asks for one specific aggregated value, which would not need a distribution of values. Thus, we made a minor wording revision to better align the statement to its intended task: "Total yearly spending rarely exceeded \$10,000". This way the question does not focus on one particular year.
- MPC.2: "The number of times that daily spending was around \$50 increased from 2020 to 2023" (cvr: 0.67, task: *Make Pattern Comparisons*). Experts' comments did not suggest particular wording changes to the T/F statement, and 4 out of 6 experts rated it a 3 or 4, so we kept it as a candidate for revision and defer to the results from test tryout (Section 6.5) for the final verdict.

The remaining seven items have a cvr of at least 80%.

5.2 Qualitative Analysis of Grading Criteria

To better understand the experts' scoring rules, and to validate (or if needed, refine) our initial clustering of answers, we asked the experts to provide free-text responses to explain their grading

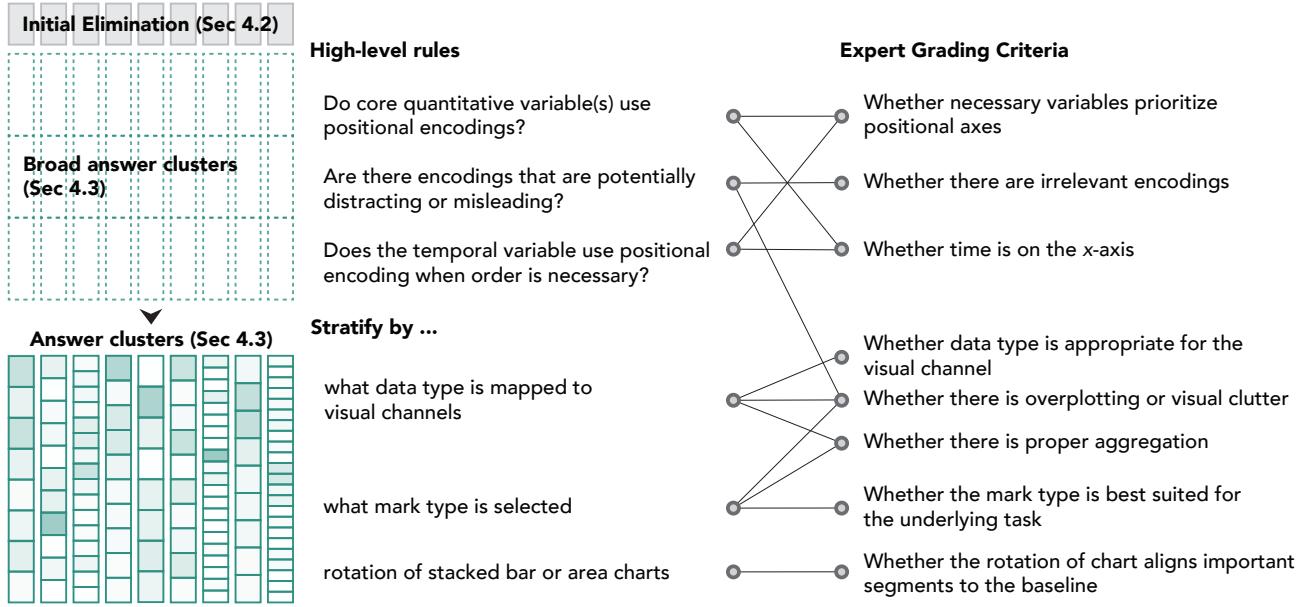


Figure 5: The mapping between our clustering rules and expert grading criteria. Each grading criterion could be mapped to one or more clustering rules we applied to cluster visualization answers of similar quality. This mapping suggests that our clusters capture the meaningful differences between answers that experts were using to form their holistic grades.

criteria for the visualization answers. We coded their responses and compared them against our clustering rules. If each expert criterion can be mapped to at least one of the clustering rules we used, then it suggests that our clusters were reasonably partitioned and captured the meaningful differences between visualization answers.

Coding Process. Two of the authors independently examined these comments and assigned codes to summarize each expert’s grading criteria. The two coders then discussed their codes for each comment, resolved differences and ambiguities, and created a final consensus codebook and consensus labels for experts’ grading criteria.⁸ Using the grading criteria codebook, we categorized the codes at the item level as item-specific rules, which we then use to assess the rules we applied for clustering visualization answers of similar quality. This ensures that our clusters are not missing some meaningful differences between answers that experts are using to form their holistic grades.

Results. We found a total of eight recurring grading criteria in the experts’ comments across all items, and each criterion can be matched to one or more rules that we first used to divide the visualization answers into clusters (see Figure 5). While some of these mappings are straightforward (e.g., *whether the rotation of chart aligns important segments to the baseline* is captured by the clustering rule *rotation of stacked bar or area charts*), a few depend on properties of our data or our tool. For example, in our datasets *visual clutter* most often occurred when a categorical variable is mapped to bar size, leading to overlapping bars of varying sizes, so the grading criterion *whether there is overplotting or visual clutter* is captured by the clustering rules *what data type is mapped to visual channels*, *what*

mark type is selected, and *are there encodings that are potentially distracting or misleading*. As explained in Section 3.3, we apply aggregation based on the selection of mark types and data variable types, so the grading criterion *whether there is proper aggregation* is captured by the clustering rules *what data type is mapped to visual channels* and *what mark type is selected*. The expert grading criteria corroborated the rules we used for clustering answers of similar quality, which suggests that our clusters do capture the meaningful differences between answers. Thus, we proceed to use these clusters and expert scores as a basis for developing an autograder to grade visualization answers (Section 5.3) and further refine the autograder with participants’ data from test tryout (Section 6.2).

5.3 Development of Initial Autograder

We develop an initial version of the autograder based on scores from the expert panel. We later refine and finalize the autograder using test tryout data in Section 6.

5.3.1 Holistic Grading. Existing approaches for grading constructed-response questions can be separated into two classes: *holistic* and *analytical* [34]. Although both approaches are based on guidelines (e.g. a rubric), the holistic approach is more top-down and aims to make a single judgement about the answer as a whole, while the analytic approach is more bottom-up, accumulating points based on specific features of the answer [34]. We used the holistic approach for grading the visualization answers, because it can better capture the *qualitative differences* between visualizations (e.g., two visualizations that both pass the same *analytical* rule of mapping necessary variables may still differ qualitatively depending on the

⁸The codebook is in supplemental materials.

chart type or the specific orientation of the chart).⁹ Thus, we asked experts to score answers on a single scale from 0 (poorly designed and unable to answer the question) to 100 (well designed and able to answer the question).

5.3.2 Grading Mechanism: Developing a Z-Score Lookup Table. To remove idiosyncrasies in experts' scoring (e.g., tendencies to use only one end of the scale), within each expert we standardized their scores into Z-scores using the mean and standard deviation of their scores. To compare the scores across items, we further standardized the expert-level Z-scores by transforming them into item-level Z-scores using the mean and standard deviation of the experts' Z-scores for each item.

We then examined each item-specific cluster of answers that experts scored. For each cluster, if the variance of the Z-scores for the answers was relatively low (meaning there are not much disagreements between experts' scores), then we assigned the mean Z-score as the final Z-score for all visualization answers in that cluster. If the variance was high (meaning there are some disagreements between experts' scores), we referred to the experts' text responses on their grading criteria as well as our own judgment to determine if the disagreement warrants (1) further dividing the cluster, (2) merging the cluster with another similar and relatively lower-variance cluster, or (3) keeping the cluster as-is and assigning the mean Z-score as the final Z-scores for the cluster. We also looked through the clusters ordered by their mean Z-scores to see if the relative order of scores reflects a reasonable ordering of the quality of the visualizations.

5.3.3 Results. After examining all 116 clusters across the 9 items, we identified and merged 2 pairs of clusters where one of the clusters within the pair had higher variance but the other cluster of the pair contained visualization answers of similar quality and had lower variance. We noticed that item MPC.2 had an especially high variance within several clusters that we believe reflected substantive disagreements between experts' scoring schemes, and we identified several instances of mean Z-scores for clusters on that item that would imply a relative ordering of answer quality that we disagreed with. We decided to keep MPC.2 provisionally for test tryout and re-examine this item after IRT analysis in Section 6.

Our initial autograder contains 114 clusters,¹⁰ each with its own corresponding Z-score, which becomes the lookup table for automatically assigning scores to visualization answers. We further refine the autograder after collecting test tryout data, described next.

6 Test Tryout

We tried out our 9 test items to refine the autograder and conduct IRT analysis. We then used the item analysis results along with evidence from Section 5 to finalize the item bank.

⁹We originally planned to grade each visualization answer holistically and analytically, but opted to just holistically because we reviewed the cluster-specific scores (Section 5.3.2) and found the holistic scores experts gave already took the important properties of the question, underlying task, and data into consideration.

¹⁰After we conducted item analysis using test tryout data, we decided not to include MPC.2 in the final bank. Therefore, there is a total of 103 clusters with their associated Z-scores in the lookup table after removing clusters corresponding to MPC.2.

6.1 Participants and Procedure

We recruited 100 participants from a Prolific pool with people who are aged 18 to 65, are fluent in English, and have normal or corrected-to-normal vision. The study was expected to take 30 minutes, and participants were compensated 6 USD for successfully completing the study. We filtered out three participants who did not complete the study and two participants who failed all of the attention check questions. The 95 remaining participants consisted of 50 males and 45 females aged 18 to 54.

At the beginning of the survey, we presented the consent form to all participants, which described what to expect in the study. The study contained two sections: the training section and the main section.¹¹ The training section included five training questions which asked participants to reproduce charts using the assessment tool. Participants must correctly answer all training questions in order to proceed. Afterwards, participants were asked to fill out the System Usability Scale (sus) [10]. We used the sus to determine if the assessment tool is easy to use (**DC.2**), which helps ensure that differences in test-takers' scores are due to their ability and not the usability of the tool.

The main section included all 9 items from our item bank and 2 attention check questions. Item order was randomized. For each item, participants were asked to create the best chart for determining whether a T/F statement is true. Then, they were asked to use the chart they created to select an answer for the T/F question and provide their reasoning. The participants were instructed that they could not go back once they proceeded to the next item. We designed the attention checks to ensure there was no room for misinterpretation if the participant had paid attention, following the format of Instructional Manipulation Checks (IMCs) [40]. Each attention check question explicitly instructed participants to select a specific mark type, drop-down value(s), and T/F answer.¹²

6.2 Autograder Refinement and Final Scores

We assigned initial grades to participants' answers using the Z-score lookup table from Section 5.3.3. That is, our autograder assigned each answer the Z-score from its corresponding cluster. If a visualization answer was eliminated during the initial elimination phase (Section 4.2), the autograder assigned it a score of -2, which is less than the minimum autograder score (-1.22) and is (by definition) two standard deviations below the mean, indicating a visualization of poor quality. Using -2 as the lowest autograder score also helps make the score distribution roughly symmetrical (the highest autograder score is 1.81).

Next, to refine the autograder and generate final grades for participants' answers, we examined each visualization answer that received a Z-score of -2 to identify any question-specific edge cases that should instead be merged into one of the existing clusters. As a result, we merged 8 types of edge cases into another cluster:

- One edge case was associated with FTC.1, which involved a rule that governed mapping quantitative variables to the color channel with mark *area*. This usually caused the visualization specifications to be unrenderable on the interface due

¹¹The demo video of the assessment tool participants used and a live demo link are both documented here: <https://osf.io/hg7kx/>.

¹²See supplemental materials for the exact attention check questions.

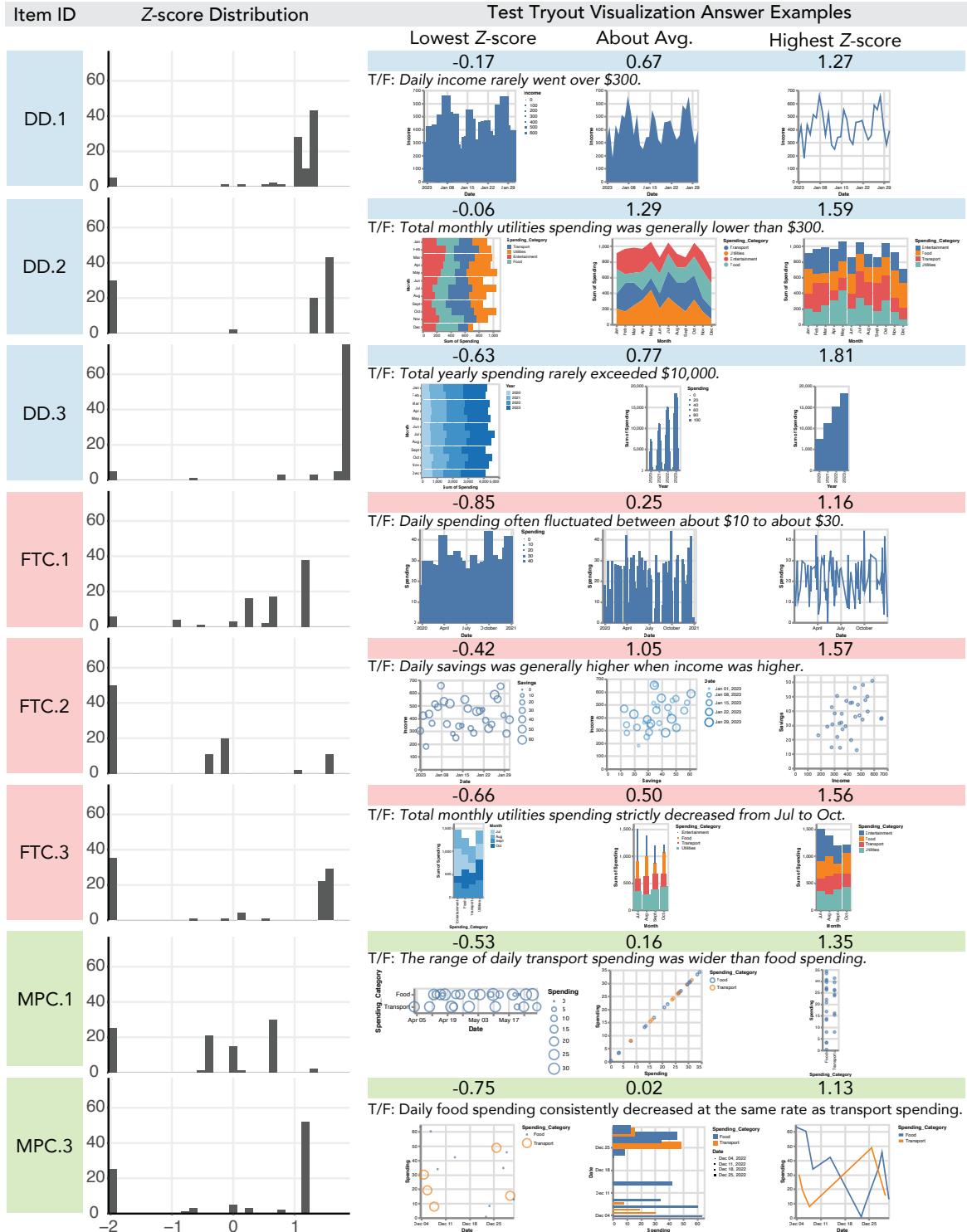


Figure 6: Distributions of Z-scores of participants' answers alongside examples of the lowest-scoring, about average, and highest-scoring visualization answers from test tryout for each item in the final bank. The lowest-scoring examples are the lowest test tryout Z-scores greater than -2.

to having too many colors. FTC.1, however, used a smaller dataset that made it an exception to this rule.

- Two cases were associated with MPC.1. In one case, the necessity of data variables varied depending on mark type (“Date” became a necessary variable along with “Spending” and “Spending Category” when the mark type is *line*, because without “Date” the line would inappropriately connect the data points in each spending category). In the other case, a visualization answer had been eliminated based on the linter rule of *Use different fields for x-axis and y-axis*, which actually did not result in a poor visualization because the task for MPC.1 (making spending range comparisons) is still doable if “Spending” was mapped to both positional axes with mark *point*. This suggests that visualization linters must be used with caution and the relevance of the linter rules may vary depending on the context and question of interest. See supplemental materials for the visualization examples.

The other 5 out of 8 edge cases are associated with MPC.2.¹³ After the autograder refinement, we reran the autograder to assign final Z-score grades to the visualization answers from test tryout. Figure 6 shows each item in the final bank alongside distributions of Z-scores of participants’ answers and example visualization answers of varying quality from test tryout.

6.3 Bayesian IRT Model

Model Specifications. The 2-parameter IRT model has been used by visualization literacy test developers to infer item easiness (i.e., how easy it is to correctly answer the item) and item discrimination (i.e., how well the item can differentiate people of different abilities) of test items [20, 24]. The traditional 2-parameter IRT model uses logistic regression, because it assumes that an answer is graded in a binary way: correct or incorrect. Since our items are graded on a continuous scale, we modified the traditional IRT model to instead use linear regression by changing the response distribution to a Gaussian. We implemented a Bayesian IRT model with the brms R package [12].

The outputs of this model are posterior distributions of item easiness, item standard deviation, and the ability of each test-taker. In the context of this model, item discrimination can be understood as the inverse of the item standard deviation: when the standard deviation of an item is high, then it cannot easily differentiate people with different abilities, because there is a lot of variance in people’s scores on that item. Thus, the higher the standard deviation, the lower the discriminating power of an item. We set the priors for the intercept to $\mathcal{N}(0, 1)$, standard deviations of ability, easiness, and responses to $\mathcal{N}^+(0, 1)$, and the correlation matrix to LKJ(2).¹⁴

Sample Size. To determine the sample size for test tryout, we conducted a pilot study ($N = 10$) to fit our modified IRT model, and used the model to simulate various sample sizes. We found a sample size of 100 was sufficient for model convergence, had reasonable posterior predictive checks, and yielded intervals for the item easiness and standard deviation parameters from the model

¹³Because we ultimately remove MPC.2 from the final item bank during test revision (Section 6.6), we defer details on its edge cases to supplemental materials.

¹⁴This model was preregistered on OSF: <https://osf.io/hg7kx/>.

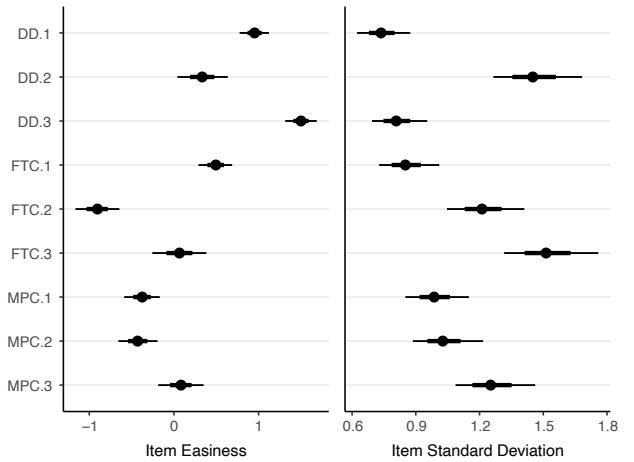


Figure 7: The item easiness estimates and item standard deviation estimates for each item. The dots in the item easiness and standard deviation coefficient plots represent the median estimates.

output that are of similar width to prior work that has also applied IRT [24].¹⁵

6.4 Descriptive Statistics and sus Score

The total completion time in minutes ranged from 8.50 to 64.76 ($M = 21.70$, $SD = 12.61$), suggesting that 30 minutes are reasonable to complete the test. The item-level correctness of the T/F questions ranged from 0.67 to 0.96 ($M = 0.82$, $SD = 0.10$). The participant-level correctness of the T/F questions ranged from 0.44 to 1 ($M = 0.82$, $SD = 0.15$).

We computed the sus scores for each participant following the established procedure [10]. The average sus score of participants is 77.76 ($SD = 19.80$), which is considered to be in the acceptable range [5]. This provides some evidence that we were able to create an interface that is easy to use (DC.2), such that unfamiliarity with the interface should be less likely to interfere with measurement of participants’ visualization encoding abilities.

6.5 IRT Analysis Results

We ran our IRT model on the test tryout data with final grades of participants’ answers for the items. The minimum bulk effective sample size is 6,530 and the minimum tail effective sample size is 6,796, and all \hat{R} values are approximately 1.

Figure 7 shows the coefficient plots displaying both item easiness and standard deviation (higher values correspond to lower discriminating power), each with 95% and 66% credible intervals (CI) and a dot indicating the median estimate. The median estimates of the item easiness parameter ranged from -0.90 to 1.50, with an average of 0.19. The median estimates of the item standard deviation parameter ranged from 0.74 to 1.51, with an average of 1.09. Figure 8 shows the distribution of the estimated abilities of participants.

¹⁵See OSF preregistration: <https://osf.io/hg7kx/>.

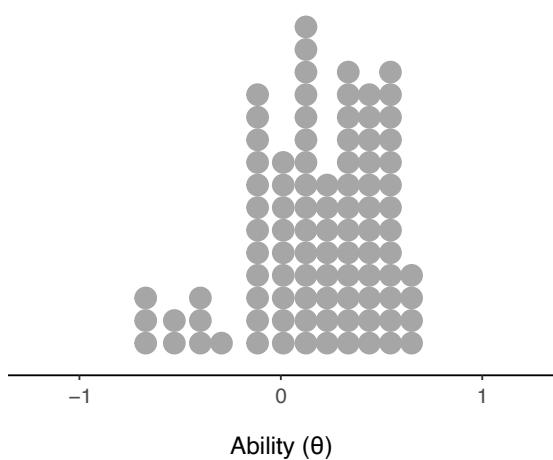


Figure 8: The distribution of the estimated abilities of participants from test tryout.

6.6 Final Item Bank

We made a holistic judgment when finalizing the item bank by considering many factors [17], including the content validity of an item, how difficult an item is, its standard deviation compared to other items in the bank, and preserving the diversity of the bank. We re-evaluated the two items identified as revision candidates ($cvi < 78\%$) in Section 5.1:

- DD.3: we retained the revised item because it has a median easiness of around 1.50, which contributes to the coverage of a wide range of abilities.
- MPC.2: we removed MPC.2 because it had a low cvi of 0.67, high variance of expert scores within answer clusters, and there is one item in the same task category (MPC.1) that has similar easiness and standard deviation, so removing MPC.2 would not necessarily affect the diversity of the item bank.

The remaining 7 items were retained because they possess at least one of the following qualities: (1) high cvi , (2) low item standard deviation, or (3) they contribute to the diversity of the item bank. We also calculated McDonald's Omega [18, 21, 44], a reliability measure widely used to assess the internal consistency of a test, and found the overall reliability ω_t to be 0.73, which is considered to be acceptable [39]. As a result, our final bank consists of 8 items, which can serve as a valid assessment of the visual encoding ability in visualization construction, and the items have varying easiness and can differentiate participants of different levels of ability.

7 Discussion

7.1 Usage Recommendations

Based on our test tryout data, the mean completion time was around 21 minutes (Section 6.4). If test administrators prefer a shorter test, they could pick from the item bank based on their desired item easiness and standard deviation estimates. For instance, if the target test population is people with lower abilities, administrators could select a set of items with low item standard deviation at

lower-ability ranges. Or if they would like to focus on a specific visualization task, such as *Describe Distributions*, then they could include only the items pertaining to that task. Items with low standard deviation (corresponding to high discriminating power in traditional IRT) should generally be preferred if one were to select a customized subset, as with prior work on assessment development using IRT [24].

Our procedure of developing AVEC could also be extended for test administrators interested in having a larger item bank. The design space we used to generate the T/F statements contains redundancies that we removed during item generation, such as T/F statements that only differ in the quantitative variable of interest (Section 3.2). One could start with the validated items in AVEC then identify other items that might have similar answer clusters as a way to expand to a larger bank. If answers for the new item fit into existing clusters, the autograder can automatically grade the answer. Otherwise, test administrators should reassess the set of visualization answers for the new item(s) and devise a scoring scheme most appropriate for the context of use.

7.2 Applications of AVEC in Future Research

One obvious area of application for AVEC is facilitating the study of the abilities involved in visualization literacy, such as the interplay between visualization interpretation and construction. For instance, AVEC could be used to investigate whether people who can appropriately map data to visual channels are also better at interpreting visualizations, and vice versa. Additionally, future work can study how much the ability to reason about potentially misleading visualizations [24] might contribute to construction-related abilities such as the visual encoding ability, which can further inform visualization teaching curricula and the design of interventions for improving a person's overall visualization literacy.

Beyond assessing a person's visual encoding ability, AVEC can also be used to evaluate the effectiveness of interventions designed to improve this ability, such as different curricula in visualization classes, through pre-post testing. These studies might help teachers identify areas that need special attention during instruction. Future work could also explore the potential for AVEC as an educational intervention itself and investigate whether providing automated and personalized feedback to users through the web-based assessment tool could improve their visual encoding ability.

7.3 Alternative Scoring Strategies

Unlike multiple-choice items, answer correctness on a constructed-response item is often not binary, and the degree of answer correctness may vary depending on the rubric or scoring scheme. We used item-specific scores from an expert panel to create a lookup table of Z -scores. The lookup table offers a flexible way of scoring the possible visualization answers in AVEC and could potentially be adjusted based on test administrators' needs. For instance, test administrators could reduce the granularity of the grades by merging clusters and their mean Z -scores, or could transform the Z -scores into simpler rankings (e.g., whole number scores from 0 to 4).

We used a holistic grading approach, as explained in Section 5.3.1, but another approach, such as analytical grading, may lead to alternate rubrics and interpretation of the scores. Alternative ways

of grading visualization answers might start with the set of expert grading criteria we summarized from the expert panel and assign points based on the number of criteria a visualization answer satisfied. Or one could take a mixed approach to grading that involves both holistically evaluating the visualization answer then applying analytical points, or vice versa. Our answer clusters and the Z-score lookup table can be easily adapted to support such alternative scoring schemes. For example, a test administrator who adopts a partially automated scoring system might use the Z-score lookup table to first filter out visualization answers that did not pass a predetermined threshold (e.g., filtering out answers that are below average), then apply analytical rules to grade the remaining visualization answers, which should greatly reduce the time for grading. Alternatively, future work could further refine the autograder using the data we collected from the expert panel. One could adapt existing methods of ranking visualizations (e.g., Draco [38, 55], MultiVision [54]) to incorporate our item-specific expert knowledge, e.g. by tailoring the knowledge base and fine-tuning the constraint weights [55] or learning weights for soft constraints from ranked pairs of visualizations (as in Draco-Learn [38]). If test administrators alter grading schemes, they should rerun the IRT analysis with the new grades, because the grades of the visualization answers would affect the item parameter estimates.

7.4 Limitations and Future Work

Individual Differences in Visualization Skills. In the test tryout study, we collected participants' basic demographics (Section 6.1) but did not collect their visualization expertise levels or familiarity with different chart types. We therefore did not explore individual differences in visualization skills or the relationships between participants' prior expertise levels and their visual encoding ability. Similar to prior work on visualization literacy assessments (e.g., Boy et al. [9], VLAT [32], CALVI [24]), this limitation does not necessarily affect AVEC's validity, which we demonstrated through having (1) items with high cv1 ratings from experts (Section 5.1) and (2) an assessment tool that preserves the combinatorial nature of selecting appropriate visual encodings while still being easy to use (Section 6.4). Exploring the relationships between different visualization skills (e.g., reading, constructing) would be interesting directions for future research.

Visualization Construction Beyond Visual Mappings. Visualization is a complex, multi-step process. Although AVEC focuses on the core of the visualization reference model (visual mappings) [13] and includes two types of aggregations (count and sum), there are many other facets of visualization construction that are important to study and measure. For instance, effectively visualizing data also requires creators to correctly handle raw data and perform appropriate data transformations (e.g., avoiding cherry picking [24, 25, 33, 35], normalizing data correctly [25, 35, 37]) for the resulting visualization to convey a clear—and correct—message. Creating effective, engaging visual narratives often requires creativity and extensive iteration (e.g., [7, 47, 48]). Expert designers may make use of interactivity or animation (e.g., [45]) to pull readers in or better convey their intended message; these techniques require time and skill to master. Future work should expand our efforts in

assessing construction-related abilities and deepen the study of the complex skills related to visualization construction.

7.5 Assessment Tools for Complex Visualization Skills

An immediate challenge facing studies of more complex skills in visualization construction is how to effectively measure such abilities. However, unless people are hand-drawing visualizations, their experience with the tool they use to construct a visualization inevitably affects the kinds of visualizations they are able to create. This presents a challenge when the goal is to measure their ability to create visualizations, not their familiarity with a specific tool. For example, a tool with a steeper learning curve that is familiar only to some users (e.g., Tableau, PowerBI) might artificially deflate unfamiliar participants' scores, even if they do have high visualization construction abilities. Thus, part of the validity of a test of construction-related abilities is dependent on demonstrating that the usability of the assessment tool does not adversely affect measurement of the ability. At a minimum, assessment tools should adhere to acceptable standards of usability (e.g., [10]) and provide adequate training to ensure participants familiarize themselves with the tool prior to assessment.

The assessment tool used to administer AVEC could be expanded to create a more versatile tool for measuring other construction-related abilities. However, arriving at a simple-to-use tool that can also measure complex construction-related abilities is a challenging design problem: higher-level skills often require more functionality, but a more complex tool runs the risk of confounding measurements of a person's ability with measurements of their familiarity with the assessment tool (or other tools like it). Future work for assessing and studying more complex skills should especially pay attention to the tool-building component and might also investigate people's ability to construct different types of visualizations with a minimal usage of computerized tools (e.g., drawing). This type of comparison study could offer insights on just how much a tool might contribute to a person's overall ability of constructing visualizations.

8 Conclusion

We systematically developed AVEC, an assessment of a person's *visual encoding ability* in visualization construction, which measures the ability to appropriately map data to visual channel(s) for effectively answering data-relevant questions. AVEC consists of 8 constructed-response items and is administered through a web-based assessment tool we built to support the combinatorial nature of selecting appropriate visual encodings. We developed an autograder for AVEC by clustering the complete set of possible visualization answers across the 9 initial items in the bank. Based on the results from an expert panel ($N = 17$), we validated our answer clusters and created a lookup table of cluster-specific Z-scores, which the autograder uses to automatically assign scores to visualization answers. The autograder significantly reduces the burden for grading constructed-response items. We refined and finalized the autograder and item bank through IRT analysis of test tryout data with 95 participants. We estimated the item easiness and standard deviation parameters for each item using IRT, which

can be used to customize the assessment based on test administrators' needs. AVEC demonstrates the importance of going beyond traditional ways of measuring visualization literacy, while shifting the focus to the more complex and higher-level construction component of visualization literacy.

Acknowledgments

Special thanks to Priyanka Nanayakkara, Sheng Long, Abhraneel Sarma, Krisha Mehta, and other members of the MU Collective at Northwestern for their valuable feedback. We extend our gratitude to our expert panel and participants for their time, as well as the anonymous reviewers for their helpful comments. The following statements are included by Ge, in accordance with the NSF Graduate Research Fellowship Program Administrative Guide (NSF 24-090): This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2234667. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- [1] Eytan Adar and Elsie Lee-Robbins. 2023. Roboviz: A Game-Centered Project for Information Visualization Education. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 268–277. doi:10.1109/TVCG.2022.3209402
- [2] Patrick Adelberger, Oleg Lesota, Klaus Eckelt, Markus Schedl, and Marc Streit. 2024. Iguanodon: A Code-Breaking Game for Improving Visualization Construction Literacy. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–14. doi:10.1109/TVCG.2024.3468948
- [3] Basak Alper, Nathalie Henry Riche, Fanny Chevalier, Jeremy Boy, and Metin Sezgin. 2017. Visualization Literacy at Elementary School. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 5485–5497. doi:10.1145/3025453.3025877
- [4] Benjamin Bach, Samuel Huron, Uta Hinrichs, Jonathan C. Roberts, and Sheelagh Carpendale. 2021. Special Issue on Visualization Teaching and Literacy. *IEEE Computer Graphics and Applications* 41, 6 (2021), 13–14. doi:10.1109/MCG.2021.3117412
- [5] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. *J. Usability Studies* 4, 3 (May 2009), 114–123.
- [6] Randy Elliot Bennett. 1991. ON THE MEANINGS OF CONSTRUCTED RESPONSE. *ETS Research Report Series* 1991, 2 (1991), i–46. doi:10.1002/j.2333-8504.1991.tb01429.x
- [7] Alex Bigelow, Steven Drucker, Danyel Fisher, and Miriah Meyer. 2017. Iterating between Tools to Create and Edit Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 481–490. doi:10.1109/TVCG.2016.2598609
- [8] Fearn Bishop, Johannes Zagermann, Ulrike Pfeil, Gemma Sanderson, Harald Reiterer, and Uta Hinrichs. 2020. Construct-A-Vis: Exploring the Free-Form Visualization Processes of Children. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 451–460. doi:10.1109/TVCG.2019.2934804
- [9] Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean-Daniel Fekete. 2014. A Principled Way of Assessing Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1963–1972. doi:10.1109/TVCG.2014.2346984
- [10] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
- [11] Katy Börner, Andreas Bueckle, and Michael Ginda. 2019. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1857–1864. doi:10.1073/pnas.1807180116
- [12] Paul-Christian Bürkner. 2021. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software* 100, 5 (2021), 1–54. doi:10.18637/jss.v100.i05
- [13] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman (Eds.). 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [14] Marianne Sheelagh Therese Carpendale. 1999. *A framework for elastic presentation space*. Ph.D. Dissertation. Simon Fraser University, CAN. Advisor(s) Fracchia, F. David. AAINQ51848.
- [15] Qing Chen, Fuling Sun, Xinyue Xu, Zui Chen, Jiazhe Wang, and Nan Cao. 2022. VizLinter: A Linter and Fixer Framework for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 206–216. doi:10.1109/TVCG.2021.3114804
- [16] Ed Huai-Hsin Chi and J.T. Riedl. 1998. An operator interaction framework for visualization systems. In *Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258)*. IEEE, Research Triangle, CA, USA, 63–70. doi:10.1109/INFVIS.1998.729560
- [17] Ronald Jay Cohen, W. Joel Schneider, and Renée Tobin. 2022. *Psychological Testing and Assessment* (10th ed.). McGraw Hill LLC, New York, NY.
- [18] Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 3 (1951), 297–334. doi:10.1007/BF02310555
- [19] Yuan Cui, Lily W. Ge, Yiren Ding, Lane Harrison, Fumeng Yang, and Matthew Kay. 2025. Promises and Pitfalls: Using Large Language Models to Generate Visualization Items. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 1094–1104. doi:10.1109/TVCG.2024.3456309
- [20] Yuan Cui, Lily W. Ge, Yiren Ding, Fumeng Yang, Lane Harrison, and Matthew Kay. 2024. Adaptive Assessment of Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2024), 628–637. doi:10.1109/TVCG.2023.3327165
- [21] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 3 (2014), 399–412. doi:10.1111/bjop.12046
- [22] FiveThirtyEight. 2022. Data Visualization – FiveThirtyEight. <https://fivethirtyeight.com/tag/data-visualization/>.
- [23] Johannes Gäbler, Christoph Winkler, Nóra Lengyel, Wolfgang Aigner, Christina Stoiber, Günter Wallner, and Simone Krigstein. 2019. Diagram Safari: A Visualization Literacy Game for Young Children. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (Barcelona, Spain) (CHI PLAY '19 Extended Abstracts). Association for Computing Machinery, New York, NY, USA, 389–396. doi:10.1145/3341215.3356283
- [24] Lily W. Ge, Yuan Cui, and Matthew Kay. 2023. CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 815, 18 pages. doi:10.1145/3544548.3581406
- [25] Lily W. Ge, Matthew Easterday, Matthew Kay, Evangelia Dimara, Peter Cheng, and Steven L Franconeri. 2024. V-FRAMER: Visualization Framework for Mitigating Reasoning Errors in Public Policy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 390, 15 pages. doi:10.1145/3613904.3642750
- [26] Lily W. Ge, Maryam Hedayati, Yuan Cui, Yiren Ding, Karen Bonilla, Alark Joshi, Alvvita Ottley, Benjamin Bach, Bum Chul Kwon, David N. Rapp, Evan Peck, Lace M. Padilla, Michael Correll, Michelle A. Borkin, Lane Harrison, and Matthew Kay. 2024. Toward a More Comprehensive Understanding of Visualization Literacy. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 494, 7 pages. doi:10.1145/3613905.3636289
- [27] Shiqing He and Eytan Adar. 2017. VizltCards: A Card-Based Toolkit for Infovis Design Education. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 561–570. doi:10.1109/TVCG.2016.2599338
- [28] Dandan Huang, Melanie Tory, Bon Adriel Aseniero, Lyn Bartram, Scott Bateman, Sheelagh Carpendale, Anthony Tang, and Robert Woodbury. 2015. Personal Visualization and Personal Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 21, 3 (2015), 420–433. doi:10.1109/TVCG.2014.2359887
- [29] Samuel Huron, Sheelagh Carpendale, Alice Thudt, Anthony Tang, and Michael Maurer. 2014. Constructive visualization. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) (DIS '14). Association for Computing Machinery, New York, NY, USA, 433–442. doi:10.1145/2598510.2598566
- [30] Samuel Huron, Yvonne Jansen, and Sheelagh Carpendale. 2014. Constructing Visual Representations: Investigating the Use of Tangible Tokens. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2102–2111. doi:10.1109/TVCG.2014.2346292
- [31] Yvonne Jansen and Pierre Dragicevic. 2013. An Interaction Model for Visualizations Beyond The Desktop. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2396–2405. doi:10.1109/TVCG.2013.134
- [32] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2017. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 551–560. doi:10.1109/TVCG.2016.2598920
- [33] Maxim Lismic, Cole Polychronis, Alexander Lex, and Marina Kogan. 2023. Misleading Beyond Visual Tricks: How People Actually Lie with Charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 817, 21 pages. doi:10.1145/3544548.3580910

- [34] Samuel A Livingston. 2009. Constructed-Response Test Questions: Why We Use Them; How We Score Them. R&D Connections. Number 11. *Educational Testing Service* (2009), 8 pages.
- [35] Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. 2022. Misinformed by Visualization: What Do We Learn From Misinformative Visualizations? *Computer Graphics Forum* 41, 3 (2022), 515–525. doi:10.1111/cgf.14559
- [36] Jock Mackinlay. 1986. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.* 5, 2 (April 1986), 110–141. doi:10.1145/22949.22950
- [37] Andrew McNutt, Gordon Kindlmann, and Michael Correll. 2020. Surfacing Visualization Mirages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376420
- [38] Dominik Moritz, Chenglong Wang, Greg L. Nelson, Haldean Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer. 2019. Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 438–448. doi:10.1109/TVCG.2018.2865240
- [39] J.C. Nunnally and I.H. Bernstein. 1994. *Psychometric Theory*. Number no. 972 in McGraw-Hill series in psychology. McGraw-Hill Companies, Incorporated, New York. <https://books.google.com/books?id=r0fuAAAAMAAJ>
- [40] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872. doi:10.1016/j.jesp.2009.03.009
- [41] Saugat Pandey and Alvitta Ottley. 2023. Mini-VLAT: A Short and Effective Measure of Visualization Literacy. *Computer Graphics Forum* 42, 3 (2023), 1–11. doi:10.1111/cgf.14809
- [42] Denise F. Polit, Cheryl Tatano Beck, and Steven V. Owen. 2007. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health* 30, 4 (2007), 459–467. doi:10.1002/nur.20199
- [43] Washington Post. 2022. Washington Post | FlowingData. <https://flowingdata.com/tag/washington-post/>.
- [44] William Revelle and David M. Condon. 2019. Reliability from α to ω : A tutorial. *Psychological Assessment* 31, 12 (2019), 1395–1411. doi:10.1037/pas0000754 Place: US Publisher: American Psychological Association.
- [45] George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko. 2008. Effectiveness of Animation in Trend Visualization. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1325–1332. doi:10.1109/TVCG.2008.125
- [46] Arvind Satyanarayanan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 341–350. doi:10.1109/TVCG.2016.2599030
- [47] Edward Segel and Jeffrey Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1139–1148. doi:10.1109/TVCG.2010.179
- [48] Randall Teal. 2010. Developing a (Non-linear) Practice of Design Thinking. *International Journal of Art & Design Education* 29, 3 (2010), 294–302. doi:10.1111/j.1476-8070.2010.01663.x
- [49] The New York Times. 2022. Graphics - The New York Times. <https://www.nytimes.com/spotlight/graphics>.
- [50] Matthew Tobiasz, Petra Isenberg, and Sheelagh Carpendale. 2009. Lark: Coordinating Co-located Collaboration with Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1065–1072. doi:10.1109/TVCG.2009.162
- [51] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. 2007. ManyEyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1121–1128. doi:10.1109/TVCG.2007.70577
- [52] Zhen Wen and Michelle Zhou. 2008. Evaluating the Use of Data Transformation for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1309–1316. doi:10.1109/TVCG.2008.129
- [53] Leland Wilkinson. 2012. *The grammar of graphics*. Springer, New York, NY. doi:10.1007/978-387-28695-0
- [54] Aoyu Wu, Yun Wang, Mengyu Zhou, Xinyi He, Haidong Zhang, Huamin Qu, and Dongmei Zhang. 2022. MultiVision: Designing Analytical Dashboards with Deep Learning Based Recommendation. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 162–172. doi:10.1109/TVCG.2021.3114826
- [55] Junran Yang, Péter Ferenc Gyarmati, Zehua Zeng, and Dominik Moritz. 2023. Draco 2: An Extensible Platform to Model Visualization Design. In *2023 IEEE Visualization and Visual Analytics (VIS)*. IEEE, Melbourne, Australia, 166–170. doi:10.1109/VIS54172.2023.00042
- [56] Caroline Ziemkiewicz and Robert Kosara. 2008. The Shaping of Information by Visual Metaphors. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1269–1276. doi:10.1109/TVCG.2008.171